Exposome, 2026, **6(1)**, osaf018
https://doi.org/10.1093/exposome/osaf018
Advance Access Publication Date: 7 January 2026
Research Article

Sociodemographic, lifestyle, environmental, and neighborhood exposures and incident type 2 diabetes in over 235,000 Dutch adults: an exposome approach

Exposome and type 2 diabetes in Dutch adults

Annelot P. Smit[1,2,*], Bette Loef[1], Jurriaan Hoekstra[1], Jeroen Lakerveld[3], Nicole A. H. Janssen[1], W. M. Monique Verschuren[1,2]

[1]National Institute for Public Health and the Environment, Bilthoven, The Netherlands
[2]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands
[3]Department of Epidemiology and Data Science, Amsterdam Public Health Research Institute, Amsterdam University Medical Center, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

*Corresponding author: Annelot P. Smit, Center for Nutrition, Prevention and Health Services, National Institute for Public Health and the Environment, P.O. Box 1, 3720 BA Bilthoven, The Netherlands (annelot.smit@rivm.nl)

Abstract

Introduction: Sociodemographic, lifestyle, environmental, and neighborhood exposures are associated with Diabetes Mellitus type 2 (DM-2). Yet, studies that include exposures from all these domains remain limited. In this study, the association between a wide range of exposures across these domains and DM-2 incidence was investigated, stratified by sex.

Methods: Data from the 2012 Dutch national health survey was used (N = 237 644), enriched with exposure data from multiple sources. In total, 57 sociodemographic, lifestyle, environmental and neighborhood exposures were included. Incidence of DM-2 was based on medication prescription from 2013 to 2022. The most important exposures for DM-2 were identified using Random Forest. Subsequently, the associations between the selected exposures and DM-2 incidence were assessed via a Cox regression, stratified by sex.

Results: In total, 5328 men and 4298 women developed DM-2 between 2013 and 2022. BMI, age and sex were identified as the most important predictors. The top 15-ranked exposures were included in the Cox regression models. BMI, age and lifestyle exposures (eg, alcohol consumption, physical activity and smoking) were associated with DM-2 in both men and women. Only in men, neighborhood exposures such as property value were associated with DM-2, while education was associated with DM-2 only in women. No associations were found for environmental exposures.

Conclusions: These results substantiate that sociodemographic and lifestyle exposures are important targets for DM-2 prevention and outperform neighborhood and environmental exposures. We observed sex-specific associations, highlighting the importance of using a sex-stratified approach in future research and clinical practice.

Key words: exposome approach; observational study; longitudinal; sex stratification; type 2 diabetes incidence.

Introduction

Diabetes Mellitus (DM) has become a major global health concern. The number of individuals suffering from DM is expected to increase to 630 million people in 2045,[1] largely due to the obesity-related increase in diabetes mellitus type 2 (DM-2).[2-4] Obesity and lack of physical activity are the most widely recognized exposures that increase the risk of DM-2.[5,6] Recently, the influence of environmental exposures on DM-2 has received increasing attention. In past studies, associations between environmental exposures (eg, air pollution) DM-2 were studied.[7-10] Higher levels of air pollution and residential noise were found to be associated with an increased risk of DM-2, while higher neighborhood walkability and green space were associated with decreased risk of DM-2.[7] All these exposures, eg, biological, lifestyle and environmental, are likely to interact with each other. Nevertheless, past studies often lacked at least one of these domains. Therefore, there is a need to examine the risk factors of DM-2 from a multi-domain exposure perspective.

A concept that takes into account this broad perspective, called the exposome, was proposed by Christopher Wild in 2005.[11] The exposome considers three categories of exposures: the internal, specific external and general external exposures. The internal exposures comprises the biological processes in the body, such as metabolic factors and circulating hormones. The specific external exposures comprises the individual-specific exposures such as lifestyle factors, medication use and occupation. The general external exposures comprises broader social

Received: August 12, 2025; **Revised:** November 20, 2025; **Accepted:** December 10, 2025

and environmental exposures related to the individual's residential context, such as population density, level of air pollution, and the amount of greenness.[11] From an analytical perspective, the exposome presents challenges; linear regression models are less suitable to take into account large numbers of exposures because that would overfit the model.[12] Combinations of conventional statistical methods like linear regression and machine learning methods could thus help to accurately capture these complex associations.

In addition, the effects of risk factors of DM-2 differ between men and women. For example, men tend to develop DM-2 at a lower body mass index (BMI) than women.[13] The differences in effects of risk factors between sexes might be primarily of a biological nature, including alterations in body composition associated with pregnancy and menopause in women.[13] Consequently, stratifying analyses by sex is essential to accurately assess the impact of risk factors on DM-2.

Studies that include a wide range of mutually adjusted exposures and investigate their association with DM-2 stratified by sex is limited. Therefore, this study applied an exposome approach by including a wide variety of exposures from the sociodemographic, lifestyle, environmental, and neighborhood domain and assessed their association with DM-2 incidence over a 10-year period in over 235 000 Dutch adults stratified by sex. We first selected the most important predictors for DM-2 via Random Forest. Subsequently, this subset was used to assess their association with DM-2 incidence in a Cox regression.

Methods

Study design

We used data from the Public Health Monitor (PHM) 2012, which is a 4-yearly cross-sectional national health survey in the Netherlands (Public Health Monitor Adults and Elderly of the Community Health Services, Statistics Netherlands and the National Institute for Public Health and the Environment, 2012, *Gezondheidsmonitor Volwassenen en Ouderen 2012, GGD'en, CBS en RIVM*). The survey is conducted by the regional Public Health Services (GGD), Statistics Netherlands (CBS) and the National Institute for Public Health and the Environment (RIVM). This survey collects information on personal and lifestyle factors, socioeconomic and health status (eg, physical, mental, and general) from the Dutch population. The PHM 2012 comprises a total of 376 384 individuals aged ≥19 years, with oversampling of individuals aged 65 years and older. All subjects gave consent for use of their data by Statistics Netherlands and RIVM. This study was conducted in accordance with the Declaration of Helsinki.

We excluded participants without a (linked) address (n = 38), with DM-2 at baseline (n = 36 432, see below "Type 2 Diabetes Mellitus (DM-2)"), who died in 2012 (shortly after filling in the questionnaire, n = 557) or had missing data on any of the exposures (n = 101 713, most missing data in questionnaire-based exposures). After these exclusions, 237 644 participants were included in this study.

Type 2 diabetes mellitus (DM-2)

To determine prevalent DM-2 cases, we used the combination of the answer on the self-reported questions "Do you have diabetes?" and "In the last 12 months, have you been treated for diabetes by a general practitioner or a specialist?" and information about diabetes medication prescriptions in 2011-2012. Medication prescription included diabetes medication with ATC codes A10A "Insulins and analogues" and A10B "Blood glucose lowering drugs, exclusive insulins". We could not distinguish between type 1 or 2 diabetes, but because of the age range of our study population (18+), we assumed that the large majority of incident cases suffered from DM-2.[14] We had no missing information on DM-2 because diabetes medication prescription is based on an external database maintained by Health Care Netherlands (Zorginstituut Nederland) and encompasses all medications covered by national obligatory basic health insurance for all residents of the Netherlands. Participants who answered "yes" to the self-reported questions and/or had prescribed diabetes medication in 2011 or 2012 were considered as prevalent cases (n = 36 432) and were excluded. We used medication prescription from 2013 to 2022 for assessing DM-2 incidence. No information about the date or dosage of prescriptions was available.

Exposures

We included a wide range of exposures which we divided into sociodemographic and lifestyle, environmental and neighborhood exposures (see Appendix 1 and Appendix 2). In general, sociodemographic and lifestyle exposures were assessed through self-report, enriched with data on standardized household income and country of origin provided by Statistics Netherlands. Data on environmental and neighborhood exposures were obtained from multiple sources; For neighborhood exposures data from the Dutch public health services and Statistics Netherlands was used. For the environmental exposures data from RIVM (air pollution, urban green spaces, noise levels) and the Geoscience and Health Cohort Consortium (obesogenic environment)[15,16] were used. The specific neighborhood and environmental exposures and their source can be found in Appendix 1. Neighborhood and environmental variables were linked to participants based on residential address.

Statistical analysis

First, Random Forest (RF) was used to identify the most important predictors of DM-2 among the different exposures. RF is a non-parametric ensemble learning method that uses a large number of decision trees[17] and was chosen for variable selection due to its ability to detect non-linear relationships and interactions, and provides robust variable importance measures even in the presence of multicollinearity. Alternative methods such as LASSO or stepwise regression were considered, but they have limitations in capturing complex patterns in large, high-dimensional epidemiological data. Furthermore, a variable importance procedure was applied for identification of individual variables that contribute the most to the model predictions. Because RF lacks the ability to model time-to-event data, individuals who were prescribed diabetes medication between 2013 and 2022 were classified as cases (independent of timing), while those who were not were classified as non-cases.

Secondly, Cox regression was used as it allows for modeling time-to-event data and this method provides hazard ratios as interpretable effect estimates. The most important predictors were selected from the RF to be included in the Cox regression models. In the Cox regression models, yearly DM-2 medication prescription was used as outcome. The Cox regression models were used to study the association of the selected exposures by RF and DM-2 incidence over 10-year period stratified by sex. All analyses were performed in R version 4.4.3.

Identifying the most important predictors

Data pre-processing

DM-2 incidence was redefined as follows for the RF models: participants were classified as a DM-2 case if they were prescribed DM-2 medication between 2013 and 2022 (n = 9626). Additionally,

follow-up duration (in years) was included as variable in the RF to account for the temporal aspect. Inclusion of highly correlated exposures could possibly influence the outcome and results of the Variable Importance (VI) ranking.[18] If the Spearman correlation was greater than 0.9 or less than −0.9, we removed the variable that was least likely to be an important risk factor for DM-2 based on our review of scientific literature. We included PM2.5 (leaving out PM10), average income of employed residents in the neighborhood (leaving out percentage of individuals with high income in the neighborhood) and NO_2 (leaving out elemental carbon (EC)). We included both the number of alcohol glasses per week and cigarettes per day, as well as the alcohol use and smoking status, since the amount and status variables can provide complementary information. Random Forest can distinguish and utilize this additional information. There was a skewed class distribution in the outcome variable of the model, because only 4.9% of all men and 3.3% of all women were DM-2 cases. We addressed this imbalance by using an undersampling approach, were we undersampled non-DM-2 cases. This method was done because undersampling is less prone to overfit the model compared to oversampling and we still had a sufficient number of observations to perform our RF models.[19] A random balanced dataset was constructed such that all DM-2 cases were included, and three times as many non-cases were randomly selected (ratio 1:3). Exact matching was not performed because the primary aim of performing the Random Forest was to select a subset of variables most important for DM-2. Demographic variables (eg, age, sex, ethnicity) were included as predictors in the models, allowing the algorithms to account for potential differences between cases and non-cases. Random undersampling preserves the overall population distribution, whereas matching could reduce data variability and model generalizability. In epidemiological studies,[20] undersampling is also known as nest case-control sampling and undersampling has been applied to several other studies.[19,21] Because only around 20% of the non-cases were selected with this undersampling method, we decided to create five random balanced datasets where each dataset consisted of a random subset of non-diabetes cases to assess an average importance ranking. A sixth random undersampled dataset was used to validate the optimal RF model with the least number of variables.

Model performance

We used the *ranger* package to construct the RF models on five undersampled datasets.[22] In these models, the parameters mtry and minimum size of the terminal nodes were optimized using the *caret* package.[23] The mtry parameter reflects how many variables are considered for each decision split and minimum size of the terminal nodes how many data points there need to be present in the terminal node to split further. The number of trees was fixed at 1000. A 5-fold cross-validation was performed to reduce the risk of overfitting and to obtain a more robust performance of the RF for each dataset. Furthermore, the receiver operator curve (ROC) and the area under the curve (AUC) of the ROC metrics were assessed via the *roc* package of the final tuned model.[24] The ROC curves were generated using the *roc* package. We also performed RF models stratified by type of exposure domain (sociodemographic, lifestyle, neighborhood, and environmental), to determine which domain contributed most to the predictive performance.[24]

Variable importance ranking

To investigate which variables were most important to predict DM-2, we computed variable Importance (VI) scores for each individual variable via the *iml* package.[25] We used a permutation approach, which quantifies the decrease in model prediction performance when a given variable is randomly permuted. For each undersampled dataset the VI procedure was performed in triplicate and the scores were averaged into a single score. This approach resulted in five average scores from the five undersampled datasets. Subsequently, we averaged these five scores into a final importance for the variable. Because the variables selected by the RF models were relatively similar for men and women, differing only in order of the same top-ranked variables, we chose to perform the RF analyses on the total study population.

Lastly, we performed a RF model to identify which subset of variables reflect the optimal RF model with the least number of variables. We only included the top 30 ranked variables, ranked by their average VI score across the five undersampled datasets, due to computational reasons. The sixth undersampled dataset was used to validate the VI ranking by observing the effect of gradually increasing the number of included variables (based on the VI order) within a RF models on the AUC. The number of variables included in the Cox regression analyses was determined by the flattening of the curve, meaning that the AUC of the selected subset of variables showed model performance almost equal to that of the full model. To ensure the inclusion of at least three exposures from each domain in the Cox regression, we additionally complemented the top three variables from each domain if needed.

Impact of identified exposures on DM-2 incidence

Cox regression analyses were performed to estimate the hazard ratios (HR) and 95% confidence intervals (CI) for the association between the exposures and DM-2 incidence. To assess sex-specific effects, all Cox regression analyses were stratified by sex. The proportional hazards assumption was checked using scaled Schoenfeld residuals. Cox regression models assume a linear relationship between the exposure and the log hazard of the outcome. Therefore, the following variables were recoded as categorical variables after the linearity check: household income (0-20; 20-40; 40-60; 60-80; 80-100 percentiles), physical activity (sex-specific quartiles), noise road (sex-specific quartiles). Moreover, glasses of alcohol per week was recoded into three categories: 0 glasses per week; 0-7 glasses per week, and >7 glasses per week for both men and women. All selected variables were only measured at baseline and were therefore time-independent in the Cox regression analysis. The time scale used in the Cox regression model was calendar years because DM-2 medication prescription was assessed yearly. Before performing the Cox regression, we verified that the correlations among all included variables were below 0.7. This criterion was met for all correlations (results not shown). Lastly, to identify interrelation between the included variables of our Cox regression models, we calculated partial correlations between either any pair of two independent continuous variables or two independent nominal variables via Spearman and Kendall correlations, conditioning for diabetes mellitus type 2 status.

Sensitivity analyses

Since especially neighborhood and environmental exposures levels (eg, $PM_{2.5}$) might change when individuals move, we performed a sensitivity analyses in which we censored all participants who moved during the period of our study. These participants were censored based on the year they moved. Furthermore, to account for potential lifestyle changes during the COVID-19 pandemic, we performed a sensitivity analyses

excluding the COVID-19 period, restricting the follow up period to 2013-2019. Lastly, to identify age-specific risk factors, we stratified our analysis by four age groups (younger than 40 years, 40-64 years, 65-79 years and 80 years and older). The variable work was recoded to having a paid job yes/no, because not all categories are applicable to each age group.

Results

Baseline characteristics

Our study population consisted of 237 644 participants of which 54% were women. Table 1 shows a selection of the included exposures by sex, with continuous variables summarized as means and standard deviations (SD) and categorical as percentages (%). Appendix 3 shows the baseline characteristics for all 57 included exposures. In total, 9626 participants developed DM-2 between 2013 and 2022, with more cases in men than in women. In addition, men were more often highly educated, more often working full-time (>32 h/wk) or retired, used more alcohol and were more often a former smoker than women. Environmental and neighborhood exposures were comparable between men and women.

Random forest—receiver operator characteristic curves (ROC)

Figure 1 shows the ROC of random forest (RF) models predicting DM-2 status in the total population. ROC curves were plotted for the separate RF models including all exposures, as well as for the subsets of sociodemographic, lifestyle, environmental, and neighborhood exposures. Based on the validation dataset, the area under the curve (AUC) value for the RF model including all exposures (0.76, 95% confidence interval (CI) = [0.76-0.77]) and the AUC of the RF model including only sociodemographic and lifestyle exposures (0.77, 95% CI = [0.77-0.78]) were similar. The AUC values of

Table 1. Selection of baseline characteristics in 2012 stratified by sex.

	Men (n = 108 528) Mean (SD)/%	**Women (n = 129 116) Mean (SD)/%**
Diabetes during follow-up, yes (%)	4.9	3.3
Sociodemographic and lifestyle		
Age (years)	55.0 (17.0)	52.4 (17.5)
Education		
Primary or less	5.8	6.0
Lower secondary	28.0	34.7
Higher secondary	31.8	30.2
University or higher	34.4	29.1
Household composition, living alone (%)	13.1	18.2
Household income [percentiles 1-100]	62.5 (25.2)	59.9 (26.3)
Work		
Retired	38.7	22.5
Employment >32 h/week	47.5	20.2
Employment 20-32 h/week	2.9	19.4
Employment 12-20 h/week	1.1	7.5
Employment <12 h/week	0.8	3.5
On disability benefits	2.7	3.4
Receiving welfare benefits	0.7	0.9
Student	3.0	4.1
Unemployed	2.1	1.9
Housemaker	0.6	16.6
Alcohol user, yes (%)	90.2	80.4
Alcohol consumption (glass/week)	9.8 (10.9)	4.9 (6.6)
Smoking status		
Current	17.8	17.0
Former	46.1	34.9
Never	36.1	48.1
Cigarettes consumption (cig/day)	2.2 (5.8)	1.8 (5.1)
Physical activity (min/week)	1083.8 (1021)	798.8 (799)
BMI (kg/m^2)	25.8 (3.4)	25.1 (4.2)
Loneliness [1-11]	2.5 (2.9)	2.5 (3.1)
Environmental		
$PM_{2.5}$	15.4 (1.5)	15.4 (1.5)
OPdtt	11.8 (2.0)	11.8 (2.0)
Noise road-traffic	50.9 (6.6)	50.9 (6.6)
Surrounding greenness	0.4 (0.1)	0.4 (0.1)
Average distance to hospital	7.3 (5.1)	7.2 (5.2)
Obesogenic environment [0-100]	19.6 (15.1)	19.4 (15.1)
Neighborhood		
Population density in neighborhood	4185.8 (3302.6)	4251.0 (3329.4)
Inhabitants between 15 and 64 in neighborhood	65.5 (6.2)	65.5 (6.3)
Non-western immigrants	9.1 (11.8)	9.2 (11.9)
Western immigrants	9.0 (4.8)	9.1 (4.8)
Financial benefits in neighborhood	70.6 (32.3)	70.8 (32.5)
Low-income households (%)	36.6 (12.6)	37.0 (12.6)
Average property value	250.9 (99.3)	250.2 (98.9)
Rental housing	38.6 (18.7)	39.2 (18.9)

Abbreviations: OPdtt, oxidate potential dithiothreitol of particle matter; PM2.5, particle matter; SD, standard deviation.

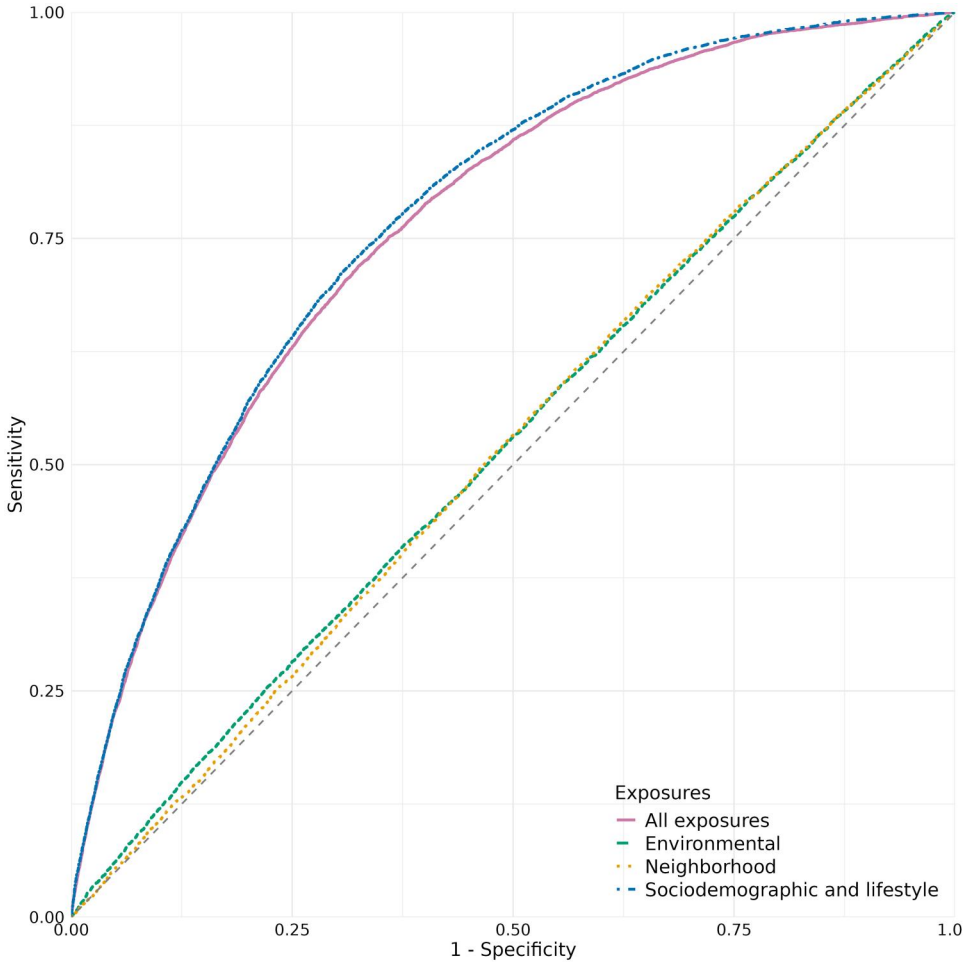

Figure 1. ROC for the RF models predicting DM-2 in the total population. ROCs are plotted for RF models including all exposures, only sociodemographic and lifestyle, only neighborhood, and only environmental exposures. The dashed grey line reflects an AUC value of 0.5.

the RF models including only environmental exposures (0.52, 95% CI = [0.52-0.53]) and only neighborhood exposures (0.52, 95% CI = [0.51-0.53]) showed low discriminatory ability and were lower than the AUC value of the RF models including all exposures and sociodemographic, lifestyle and exposures.

Identifying the most important predictors for DM-2

Figure 2 shows the average importance ratio of the top 30 exposures in predicting DM-2 status in the total population, based on the averaged results of the five undersampled datasets. The top 10-ranked exposures all belonged to the sociodemographic and lifestyle domain, with BMI and age ranked first and second. Property value in the neighborhood was the highest-ranked neighborhood-related exposure (position 14), and surrounding greenness in 100 m buffer (indicating an individuals' surrounding green space) was the highest-ranked environment-related exposure (position 11). The ranking of all 57 included exposures can be found in Appendix 4.

Figure 3 shows the increase in AUC value with an increasing number of exposures based on the sixth undersampled dataset; the validation dataset. We selected the top 15-ranked exposures because additional exposures did not add substantially to the predictive performance of the model. The top 15-ranked exposures from the importance ranking of Figure 2 were selected for our Cox regression models. To ensure the inclusion of at least three exposures from each domain in the Cox regression, "Number of households in the neighborhood" and "Average income of employed residents in the neighborhood" were also selected.

Associations of identified exposures and DM-2 incidence

The proportional hazards assumption was not violated (Appendix 5, Appendix 6). The results of the Cox regression showed that a higher BMI, a higher age (per 5 years), smoking more cigarettes per day (per 10 cigarettes), and being a former or current smoker were associated with an increased risk of DM-2 during 10-years of follow-up in both men in women (Figure 4, Appendix 7). The association of increasing BMI, age and being a former smoker were stronger in men than in women. Only in women, higher education (HRs ranging from 0.89, 95% CI = [0.80; 0.98] for lower secondary, to 0.66, 95%CI [0.57; 0.75] for university or higher) was associated with a decreased risk of DM-2. In men, work was associated with an increased risk of DM-2, whereas in women, work was associated with a decreased risk of DM-2. Being in the highest income group was associated with a decreased risk of DM-2 in both men and women. More physical activity was associated with a decreased risk of DM-2 for all quartiles in men and only for the third quartile in women. Alcohol consumption was associated with a reduced risk of

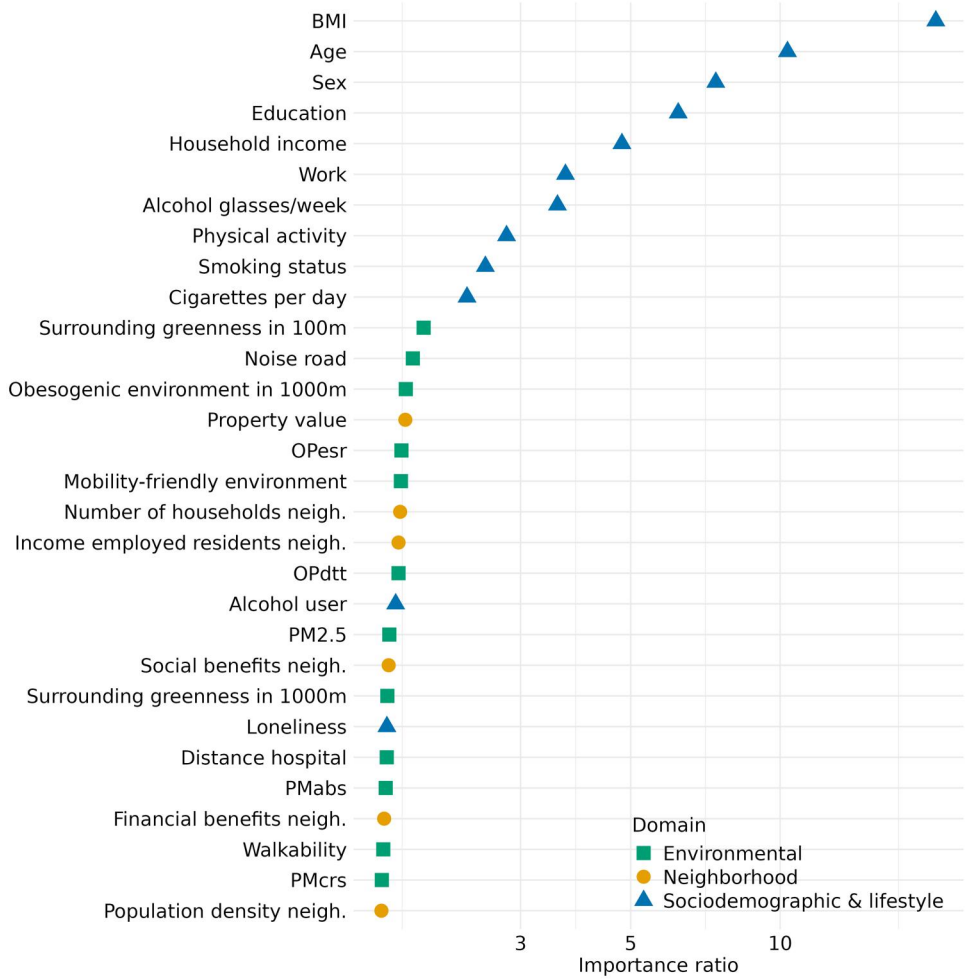

Figure 2. Variable importance ranking of the top 30 ranked exposures in the total population. The x-axis shows the importance ratio which reflects the decrease in model performance when the values of a predictor are randomly permuted. This procedure was performed in triplicate for each of the five undersampled datasets. The average importance ratio across all five undersampled datasets is shown. Color and shapes reflect the domain of the exposure. OPesr: oxidative potential electron spin resonance, neigh: neighborhood, PM2.5: Particulate matter 2.5, OPdtt: oxidate potential dithiothreitol of particle matter, PMabs: Particulate matter absorbance, PMcrs: Particulate matter coarse.

DM-2, with stronger associations in women than in men (eg, HR = 0.65, 95% CI = [0.59; 0.72] >7gl/wk in women and HR = 0.72, 95%CI [0.66; 0.78] >7gl/wk in men). Only in men, a lower property value in the neighborhood, more households in the neighborhood, and a higher average income of employed residents in the neighborhood were associated with an increased risk of DM-2.

Both sensitivity analyses yielded similar results, except that some effect estimates increased in magnitude, possibly due to the shorter follow-up period (Appendix 8 & Appendix 9). Also across age groups, no large differences were found, except that among individuals younger than 40, women had a higher risk of developing type 2 diabetes than men, whereas in the older age groups, men had a higher risk than women (Appendix 10).

Partial correlation analysis

For the significant variables from the Cox regression in men, property value with income per employed resident in the neighborhood (0.72) and property value with number of households (−0.33) had a moderate to high correlation. In women, education and income (0.28) were moderately correlated. For the other significant variables in the Cox regression of both men and women, we found low partial correlations (Appendix 10).

Discussion

This study adopted an exposome approach by incorporating a wide variety of exposures from different domains, to identify important predictors of DM-2 and to subsequently investigate their associations with 10-year DM-2 incidence. Across a large set of sociodemographic and lifestyle, environmental, and neighborhood exposures, we found that the most important predictors of DM-2 belonged to the sociodemographic and lifestyle domain, ie, BMI, age, and sex. When examining the associations of the most important exposures with DM-2 incidence, we found that a higher BMI, higher age, and lifestyle exposures as higher alcohol use, lower physical activity, and smoking had the strongest associations with an increased risk of DM-2 in both men and women. Moreover, neighborhood exposures as property value in the neighborhood and average income of employed residents in the neighborhood were associated with DM-2 only in men, while higher education was associated with DM-2 only in women.

Sociodemographic and lifestyle exposures, like BMI, age, and alcohol use were identified as the most important predictors for DM-2 and showed significant associations with DM-2 incidence in the Cox regression analyses. While these exposures have been well-studied previously,[5,6] our study provided a comparative

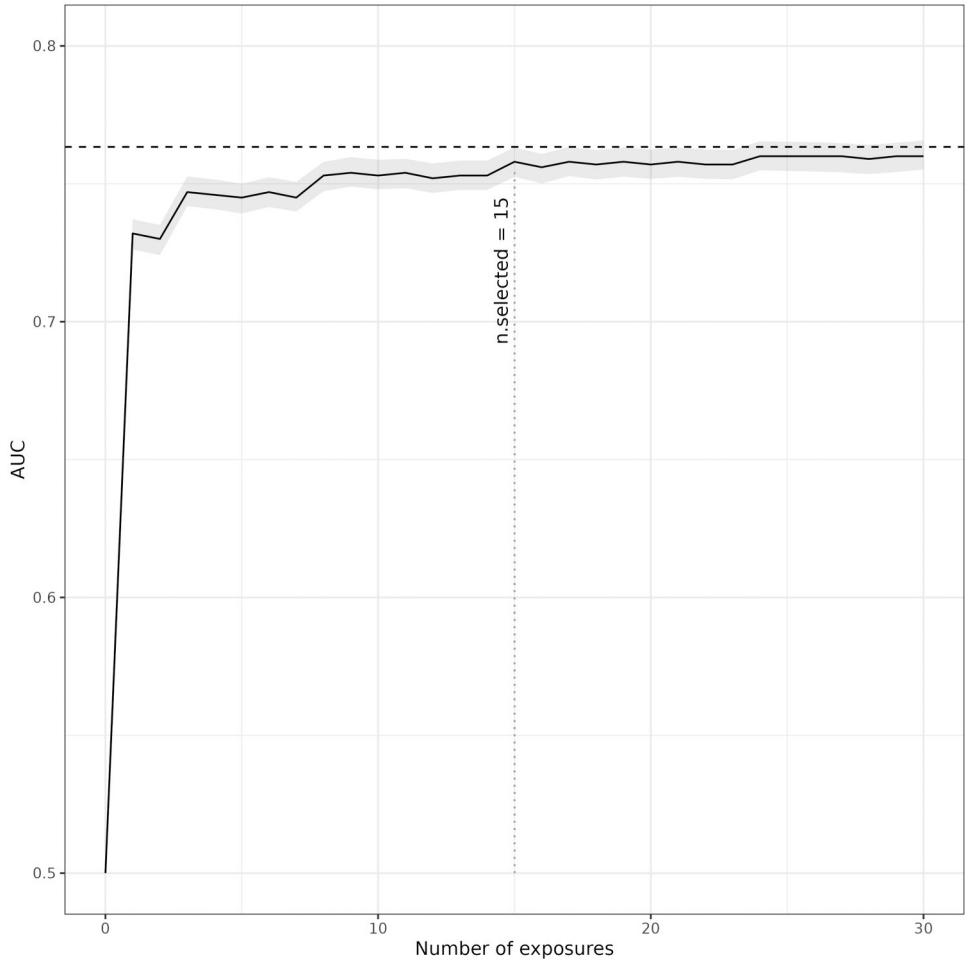

Figure 3. Validation plot on the AUC value of first 30 exposures when the number of exposures is increased in the RF models. 5-fold cross-validation was used to calculate the AUC value and its 95% CI on the sixth undersampled dataset. Variables were added sequentially based on their variable importance ranking. The horizontal line reflects the AUC of the RF model including all exposures.

ranking of their predictive performance and effect sizes when exposures from multiple domains are included. The strong association between BMI and DM-2 in our study is in line with the close alignment of BMI with internal biological processes relevant to DM-2 development.[3,4,11] Our finding thus underscores the association of BMI as major driver for DM-2. In addition, alcohol consumption was associated with a reduced risk of DM-2 compared to no alcohol consumption. While moderate alcohol consumption has shown protective effects for DM-2 in meta-analyses,[26] these findings do not imply that drinking alcohol should be encouraged.[27] An important consideration in studying the association between alcohol consumption and DM-2 is the reference group, which is often a heterogeneous mix of lifelong abstainers as well as former drinkers who quit voluntarily or for health reasons.[26] Unfortunately, we lacked information on the reasons why individuals stopped drinking alcohol.

None of the environmental exposures (ie, surrounding greenness in 100 m, obesogenic environment in 1000 m and road traffic noise) were associated with DM-2, even though emerging evidence suggests that lower levels of green and higher levels of road traffic noise are associated with an increased risk of DM-2.[7,28] With regard to road traffic noise, our finding is consistent with a previous study conducted in the same cohort.[29] As suggested in this previous study, the lack of adjustment for multiple environmental exposures in some previous studies may have led to an overestimation of the road traffic noise effect, potentially explaining the discrepancies in findings.[29] Moreover, several alternative explanations should be considered when interpreting our findings. First, some environmental exposures might be precursors of an individual's behavior. The built environment has been shown to influence DM-2 directly as well as indirectly by influencing an individual's behavior.[30] Therefore, a healthy environment might not be strongly associated with DM-2 when adjusted for lifestyle exposures, but these exposures could be an encouraging stimulus for a healthy lifestyle.[30] Secondly, environmental exposures were measured only at each individual's residential address at baseline. Because environmental concentrations vary substantially across different locations, this approach may not accurately capture actual exposure levels. Therefore, we performed sensitivity analyses in which we censored participants at the year they moved to account for large changes in environmental exposures levels. This sensitivity analyses yielded similar results as our main analysis. Lastly, environmental and neighborhood exposures have been shown to exhibit a more modest effect in complex multivariate models,[8] suggesting that these exposures lack substantial predictive value, especially when sociodemographic and lifestyle exposures are included. Our results do therefore not imply that environmental and neighborhood exposures not included in the Cox regression, are not associated with DM-2 incidence.[31] For example, although

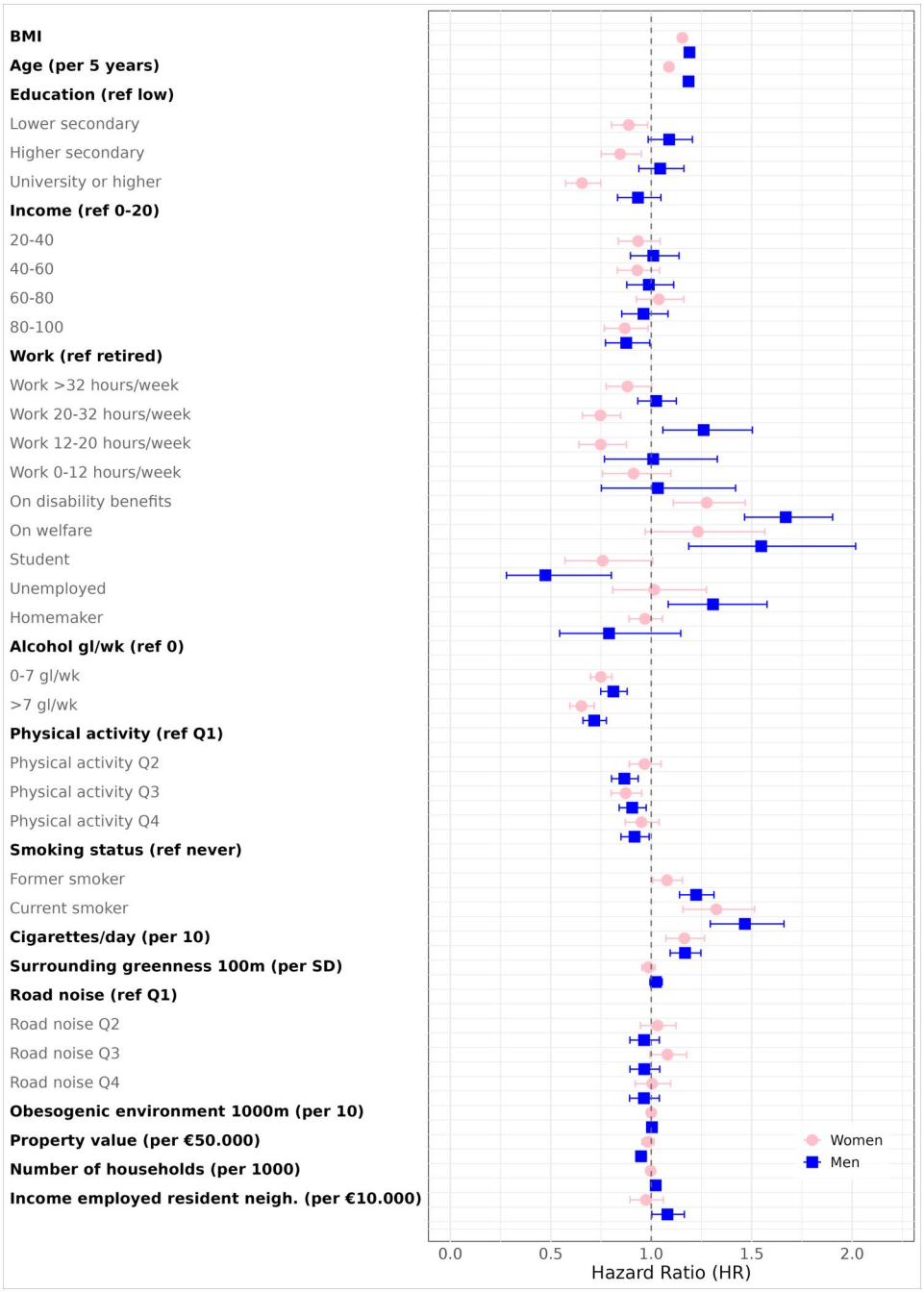

Figure 4. Forest plot of the hazard ratios (HRs) and 95% confidence intervals (CIs) for the association between the selected exposures by RF and DM-2 incidence. Results are shown separately for women (n = 129 116, pink circles) and men (n = 108 528, blue squares). Ref: reference group. Neigh.: neighborhood. Physical activity Q1: men <360 min/week, women <240 min/week, Q2: men 360-750 min/week, women 240-540 min/week, Q3: men 750-1500 min/week, women 540-1080 min/week, Q4: men >1500, women >1080 min/week. Road noise Q1: men <46.3 dB, women <46.3 dB, Q2: men 46.3-50.1 dB, women 46.3-50.1 dB, Q3: 50.1, 54.8 dB, women 50.1-54.9 dB, Q4: men >54.8 dB, women >54.9 dB.

our results showed that number of households did not have a high predictive value for DM-2, it did have a statistically significant association with DM-2 risk in men. Nevertheless, by evaluating a wide variety of exposures simultaneously, our study suggests that sociodemographic and lifestyle-related exposures have stronger predictive values and associations with DM-2 risk than environmental or neighborhood exposures.

This study also highlighted that effect sizes of some exposures differed between men and women. First, the effect size of age was stronger in men than in women. This finding is in line with previous research that showed that men are on average diagnosed with DM-2 at a younger age than women.[32] Secondly, higher education and working had a protective effect on DM-2 incidence in women, but not in men. The protective association between education and DM-2 risk in women has also been found previously.[13,33,34] However, research investigating the mechanisms behind this association is lacking.[33] Third, higher physical activity was associated with a decreased risk across all quartiles in men, but only in the third quartile among women. In our study, physical activity included both leisure time as well as occupational time physical activity. In a past study that also included occupational time in physical activity, it has been

suggested that including this type of physical activity gave a more reduced risk in men than in women, which is in line with our results.[35] This discrepancy might be explained because men are more likely to perform physically demanding tasks, even within the same occupations as women, possibly due to differences in physical strength.[36] Moreover, in a past study, it has been shown that men have more sedentary behavior than women across the day, but when men are physically active, the intensity of their activity is higher. In contrast, women are moderately active across the whole day. As a result, men might accumulate a higher total amount of metabolic equivalents (METs) than women.[37] Fourth, only in men, a lower property value, higher number of households, and higher average income of employed residents in neighborhood were associated with higher risk of DM-2. A lower average property value in the neighborhood was also associated with risk of DM-2 in previous studies, however these results were not stratified by sex.[8,38] The difference in average income of employed residents is in line with previous research that also showed stronger effects of this exposure in men than in women.[39] Nevertheless, because studies investigating sex-stratified associations between neighborhood and environmental exposures and DM-2 incidence are still limited, we recommend future studies to stratify their analyses by sex as these exposures might differently impact men and women's risk on DM-2 and these effects extend beyond sociodemographic and lifestyle exposures.

A strength of this study was the large sample size. Therefore we were able to stratify by sex and include a wide variety of exposures from different domains. In addition, because we used an external database providing yearly medication prescription to determine DM-2 incidence, we had complete follow-up data for all participants. Nevertheless, several limitations need to be acknowledged. First, based on medication prescription it could not be determined if an individual suffered from diabetes mellitus type 1 or type 2. However, as mentioned, because our study population was aged 18 years and older, it could be assumed that the large majority of incident diabetes cases will be type 2.[14] Nevertheless, type 1 diabetes or other forms of diabetes may still occur among individuals in our study. Secondly, by design, this study population consists of an oversampling of individuals aged 64 years and older. Because most individuals suffering from DM-2 receive their diagnosis between ages 45-64 years,[40] our incidence rates might have been lower. However, this type of selection bias might have led to an underestimation of our associations. Moreover, all exposures were measured only at baseline, limiting our ability to assess potential changes in exposure over time.[41] The effect sizes, of particularly lifestyle, exposures in our main model may therefore have been attenuated due to changes in participants' lifestyle behaviors over time.

In conclusion, in this study which included a wide range of exposures from different domains, BMI, age, and lifestyle exposures showed the strongest associations with DM-2 incidence. While these exposures are widely recognized risk factors for DM-2, our findings substantiate that sociodemographic and lifestyle exposures are important targets for DM-2 prevention and outperform neighborhood and environmental exposures. In addition, we observed sex-specific associations; neighborhood exposures were associated with DM-2 incidence only in men, while education was associated with DM-2 incidence only in women. These findings highlight the importance of using a sex-stratified approach in future research to better understand how risk factors for DM-2 incidence differ between men and women.

Acknowledgments

We thank all the respondents of the PHM (Gezondheidsmonitor Volwassenen GGD-en, CBS en RIVM) 2012 and 2016. The Public Health Services (GGD), Statistics Netherlands (CBS) and National Institute for Public Health and the Environment (RIVM) conducted the monitoring. CBS facilitated the statistical analyses. This research was funded by the Strategic Program RIVM (Dutch National Institute for Public Health and the Environment), grant number: S/010003/03/CP. Geo-data were collected as part of the Geoscience and Health Cohort Consortium (GECCO) on the obesogenic environment, which was financially supported by the Netherlands Organisation for Scientific Research (NWO), the Netherlands Organisation for Health Research and Development (ZonMw), and Amsterdam UMC. More information on GECCO can be found on www.gecco.nl. The authors thank A. Nicolai for her valuable insights on the statistical aspects of this research.

Author contributions

Annelot P. Smit (Conceptualization [Equal], Formal analysis [Lead], Investigation [Lead], Methodology [Lead], Visualization [Lead], Writing—original draft [Lead]), Bette Loef (Conceptualization [Supporting], Methodology [Supporting], Supervision [Lead], Writing—review & editing [Equal]), Jurriaan Hoekstra (Methodology [Equal], Writing—review & editing [Equal]), Jeroen Lakerveld (Data curation [Lead], Resources [Lead], Writing—review & editing [Supporting]), Nicole A. H. Janssen (Conceptualization [Lead], Data curation [Lead], Funding acquisition [Lead], Project administration [Supporting], Resources [Lead], Supervision [Lead], Writing—review & editing [Equal]), W. M. Monique Verschuren (Conceptualization [Lead], Funding acquisition [Lead], Methodology [Supporting], Project administration [Lead], Supervision [Lead], Writing—review & editing [Equal]).

Data availability and materials

The dataset used in this study was constructed based on multiple existing datasets from the Dutch public health services and Statistics Netherlands (public health monitor, neighborhood statistics), RIVM (air pollution, urban green buffers, noise levels) and GECCO (obesogenic environment). Access to the PHM data used in this study is restricted due to licensing agreements, and therefore the final datasets are not publicly available in an online repository. Access to the PHM data can be requested by contacting the Dutch public health services (monitorgezondheid@ggdghor.nl). After receiving permission from the Dutch public health services, the full datasets utilized in this research may be obtained from the corresponding author upon a reasonable request. The publicly available data sources referenced in this study are accessible and listed within the publication. The following open-access resources were utilized:

Statistics Netherlands, District and neighborhoods map 2012. Accessible from: https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/wijk-en-buurtkaart-2012

RIVM, Atlas Living Environment. Accessible from: https://www.atlasleefomgeving.nl/kaarten

RIVM, Atlas Natural Capital. Accessible from: https://www.atlasnatuurlijkkapitaal.nl/kaarten

Supplementary material

Supplementary material is available at *Exposome online*.

Conflicts of interest

None declared.

Funding

This research was funded by the Strategic Program RIVM (Dutch National Institute for Public Health and the Environment), grant number: S/010003/03/CP.

Declaration of use of artificial intelligence (AI)

The Artificial intelligence-assisted technologies ChatGPT (OpenAI) was used to support the development of R scripts and to improve the clarity and language of the manuscript text. All AI-generated content was critically reviewed and edited by the authors.

References

1. Standl E, Khunti K, Hansen TB, Schnell O. The global epidemics of diabetes in the 21st century: current situation and perspectives. *Eur J Prev Cardiol*. 2019;26(2_suppl):7–14. https://doi.org/10.1177/2047487319881021
2. Viner R, White B, Christie D. Type 2 diabetes in adolescents: a severe phenotype posing major clinical challenges and public health burden. *Lancet*. 2017;389(10085):2252–2260. https://doi.org/10.1016/S0140-6736(17)31371-5
3. Zheng Y, Ley SH, Hu FB. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. *Nat Rev Endocrinol*. 2018;14(2):88–98. https://doi.org/10.1038/nrendo.2017.151
4. Yu H-j, Ho M, Liu X, Yang J, Chau PH, Fong DYT. Association of weight status and the risks of diabetes in adults: a systematic review and meta-analysis of prospective cohort studies. *Int J Obes (Lond)*. 2022;46(6):1101–1113. https://doi.org/10.1038/s41366-2–01096-1
5. World Health Organization. Global report on diabetes. 2016. 2018.
6. Uusitupa M, Khan TA, Viguiliouk E, *et al.* Prevention of type 2 diabetes by lifestyle changes: a systematic review and meta-analysis. *Nutrients*. 2019;11(11):2611. https://doi.org/10.3390/nu11112611
7. Beulens JWJ, Pinho MGM, Abreu TC, *et al.* Environmental risk factors of type 2 diabetes-an exposome approach. *Diabetologia*. 2022;65(2):263–274. https://doi.org/10.1007/s00125-1–05618-w
8. Ohanyan H, Portengen L, Kaplani O, *et al.* Associations between the urban exposome and type 2 diabetes: results from penalised regression by least absolute shrinkage and selection operator and random forest models. *Environ Int*. 2022;170:107592. https://doi.org/10.1016/j.envint.2022.107592
9. Sørensen M, Poulsen AH, Hvidtfeldt UA, *et al.* Air pollution, road traffic noise and lack of greenness and risk of type 2 diabetes: a multi-exposure prospective study covering Denmark. *Environ Int*. 2022;170:107570. https://doi.org/10.1016/j.envint.2022.107570
10. Feyissa TR, Wood SM, Vakil K, *et al.* The built environment and its association with type 2 diabetes mellitus incidence: a systematic review and meta-analysis of longitudinal studies. *Soc Sci Med.Medicine* 2024;361:117372. https://doi.org/10.1016/j.socscimed.2024.117372
11. Wild CP. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomarkers Prev*. 2005;14(8):1847–1850. https://doi.org/10.1158/5–9965.Epi-5–0456
12. Isola S, Murdaca G, Brunetto S, Zumbo E, Tonacci A, Gangemi S. The use of artificial intelligence to analyze the exposome in the development of chronic diseases: a review of the current literature. *Informatics*. 2024;11(4):86. https://doi.org/10.3390/informatics11040086
13. Kautzky-Willer A, Leutner M, Harreiter J. Sex differences in type 2 diabetes. *Diabetologia*. 2023;66(6):986–1002. https://doi.org/10.1007/s00125-3–05891-x
14. Carstensen B, Rønn PF, Jørgensen ME. Prevalence, incidence and mortality of type 1 and type 2 diabetes in Denmark 6–2016. *BMJ Open Diabetes Res Care*. 2020;8(1):e001071. https://doi.org/10.1136/bmjdrc-2019-001071
15. Lakerveld J, Wagtendonk A, Vaartjes I, Karssenberg D, GECCO Consortium. Deep phenotyping meets big data: the Geoscience and hEalth Cohort COnsortium (GECCO) data to enable exposome studies in The Netherlands. *Int J Health Geogr*. 2020;19(1):49. https://doi.org/10.1186/s12942-0–00235-z
16. Timmermans EJ, Lakerveld J, Beulens JWJ, *et al.* Cohort profile: the Geoscience and Health Cohort Consortium (GECCO) in the Netherlands. *BMJ Open*. 2018;8(6):e021597. https://doi.org/10.1136/bmjopen-8–021597
17. Breiman L. Random forests. *Machine Learning*. 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324
18. Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. *BMC Bioinformatics*. 2008;9:307. https://doi.org/10.1186/1–2105-9–307
19. Dubey R, Zhou J, Wang Y, Thompson PM, Ye J. Analysis of sampling techniques for imbalanced data: an n = 648 ADNI study. *Neuroimage*. 2014;87:220–241. https://doi.org/10.1016/j.neuroimage.2013.10.005
20. Langholz B, Borgan ØR. Counter-matching: a stratified nested case-control sampling method. *Biometrika*. 1995;82(1):69–79. https://doi.org/10.1093/biomet/82.1.69
21. Pias TS, Su Y, Tang X, Wang H, Yao D. Undersampling for fairness: achieving more equitable predictions in diabetes and prediabetes. *medRxiv*. 23289405;2023(2023.05.02). https://doi.org/10.1101/2023.05.02.23289405
22. Wright MN, Ziegler A. ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J Stat Soft*. 2017;77(1):1–17. https://doi.org/10.18637/jss.v077.i01
23. Kuhn M. Building predictive models in R using the caret package. *J Stat Soft*. 2008;28(5):1–26. https://doi.org/10.18637/jss.v028.i05
24. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. *Bioinformatics*. 2005;21(20):3940–3941. https://doi.org/10.1093/bioinformatics/bti623
25. Molnar C, Casalicchio G, Bischl B. iml: an R package for interpretable machine learning. *Joss*. 2018;3(26):786. https://doi.org/10.21105/joss.00786
26. Knott C, Bell S, Britton A. Alcohol consumption and the risk of type 2 diabetes: a systematic review and dose-response meta-

analysis of more than 1.9 million individuals from 38 observational studies. *Diabetes Care*. 2015;38(9):1804–1812. https://doi.org/10.2337/dc15-0710
27. Health Council of the Netherlands. *Dutch Dietary guidelines 2015*. Vol. publication no. 2015/26e. 2015.
28. Zare Sakhvidi MJ, Zare Sakhvidi F, Mehrparvar AH, Foraster M, Dadvand P. Association between noise exposure and diabetes: a systematic review and meta-analysis. *Environ Res*. 2018;166:647–657. https://doi.org/10.1016/j.envres.2018.05.011
29. Klompmaker JO, Janssen NAH, Bloemsma LD, *et al.* Associations of combined exposures to surrounding green, air pollution, and road traffic noise with cardiometabolic diseases. *Environ Health Perspect*. 2019;127(8):87003. https://doi.org/10.1289/ehp3857
30. den Braver NR, Lakerveld J, Rutters F, Schoonmade LJ, Brug J, Beulens JWJ. Built environmental characteristics and diabetes: a systematic review and meta-analysis. *BMC Med*. 2018;16(1):12. https://doi.org/10.1186/s12916-7–0997-z
31. Poldrack RA, Huckins G, Varoquaux G. Establishment of best practices for evidence for prediction: a review. *JAMA Psychiatry*. 2020;77(5):534–540. https://doi.org/10.1001/jamapsychiatry.2019.3671
32. Cho NH, Shaw JE, Karuranga S, *et al.* IDF Diabetes Atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Res Clin Pract*. 2018;138:271–281. https://doi.org/10.1016/j.diabres.2018.02.023
33. Kautzky-Willer A, Harreiter J, Pacini G. Sex and gender differences in risk, pathophysiology and complications of type 2 diabetes mellitus. *Endocr Rev*. 2016;37(3):278–316. https://doi.org/10.1210/er.2015-1137
34. Robbins JM, Vaccarino V, Zhang H, Kasl SV. Socioeconomic status and diagnosed diabetes incidence. *Diabetes Res Clin Pract*. 2005;68(3):230–236. https://doi.org/10.1016/j.diabres.2004.09.007
35. Ekelund U, Palla L, Brage S, *et al.* Physical activity reduces the risk of incident type 2 diabetes in general and in abdominally lean and obese men and women: the EPIC-InterAct study. *Diabetologia*. 2012;55(7):1944–1952. https://doi.org/10.1007/s00125-2–2532-2
36. Hooftman WE, van der Beek AJ, Bongers PM, van Mechelen W. Gender differences in self-reported physical and psychosocial exposures in jobs with both female and male workers. *J Occup Environ Med*. 2005;47(3):244–252. https://doi.org/10.1097/01.jom.0000150387.14885.6b
37. Paraschiakos S, Bogaards FA, Knobbe A, Slagboom PE, Beekman M. Changes in the physical behaviour of older adults during the 13 weeks GOTO intervention explain a boost in immunometabolic health. *medRxiv* 23299026;2023(2023.11.26). https://doi.org/10.1101/2023.11.26.23299026
38. Consolazio D, Koster A, Sarti S, *et al.* Neighbourhood property value and type 2 diabetes mellitus in the Maastricht study: a multilevel study. *PLoS One*. 2020;15(6):e0234324. https://doi.org/10.1371/journal.pone.0234324
39. Wu H, Bragg F, Yang L, *et al.* Sex differences in the association between socioeconomic status and diabetes prevalence and incidence in China: cross-sectional and prospective studies of 0.5 million adults. *Diabetologia* 2019;62(8):1420–1429. https://doi.org/10.1007/s00125-9–4896-z
40. Zhang H-J, Feng J, Zhang X-T, Zhang H-Z. Age at type 2 diabetes diagnosis and the risk of mortality among US population. *Sci Rep*. 2024;14(1):29155. https://doi.org/10.1038/s41598-4–80790-8
41. Schermer EE, Engelfriet PM, Blokstra A, Verschuren WMM, Picavet HSJ. Healthy lifestyle over the life course: population trends and individual changes over 30 years of the Doetinchem Cohort Study. *Front Public Health*. 2022;10:966155. https://doi.org/10.3389/fpubh.2022.966155

Exposome, 2026, 6, 1–11
https://doi.org/10.1093/exposome/osaf018
Research Article